
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Drug Discovery Today • Volume xxx, Number xx • xxxx 2021 REVIEWS



# Artificial intelligence for the discovery of novel antimicrobial agents for emerging infectious diseases

Adam Bess [a], Frej Berglind [a], Supratik Mukhopadhyay [a], Michal Brylinski [a], Nicholas Griggs [b], Tiffany Cho [b], Chris Galliano [c], Kishor M. Wasan [c,d,*]

[a] Department of Computer Sciences, Louisiana State University, Baton Rouge, LA, USA
[b] Trinity Consultants Inc., Los Angles, CA, USA
[c] Skymount Medical US Inc, New Orleans, LA, USA
[d] Department of Urologic Sciences, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

**The search for effective drugs to treat new and existing diseases is a laborious one requiring a large investment of capital, resources, and time. The coronavirus 2019 (COVID-19) pandemic has been a painful reminder of the lack of development of new antimicrobial agents to treat emerging infectious diseases. Artificial intelligence (AI) and other *in silico* techniques can drive a more efficient, cost-friendly approach to drug discovery by helping move potential candidates with better clinical tolerance forward in the pipeline. Several research teams have developed successful AI platforms for hit identification, lead generation, and lead optimization. In this review, we investigate the technologies at the forefront of spearheading an AI revolution in drug discovery and pharmaceutical sciences.**

**Keywords: Artificial intelligence; Infectious diseases; Antimicrobial agents; COVID-19**

## Artificial intelligence has been transformative in several areas of human endeavor

Exponential progress in AI and its applications has occurred during the past 15 years.[1] AI-based conversational assistants are now powering consumer devices, such as Amazon's Alexa; self-driving cars have been registering hundreds of thousands of miles on American roads[2]; AI has beaten world champions in GO, chess, and other games[3–4]; AI-based systems are assisting doctors in medical diagnosis and treatment[5–8]; AI is helping map the canopy cover across the continental USA[9–11]; and a combination of AI and immersive virtual reality is assisting construction engineers to design energy-efficient buildings[12–13]. In summary, AI is influencing every aspect of human life from transportation[14] to stock trading.[15] However, the influence of AI on drug discovery and development has been minimal thus far.

## AI currently lacks impact on drug discovery

Although it is undeniable that the application of AI in pharmaceutical sciences holds tremendous promise, the current limited impact of AI on drug discovery can be attributed to multiple factors. A lack of standardized labeled benchmark data sets has been one of the major hurdles of AI-driven drug discovery. The recent AI revolution has been fueled by the availability of cheap computing power and large volumes of data that can be easily shared through the internet. For example, progress in computer vision has been dramatically accelerated by the creation of the benchmark ImageNet data set.[16] Despite several attempts, such as DrugBank,[17] BindingDB,[18] KEGG,[19] Supertarget,[20] DUD-E,[21] and others, all-encompassing benchmark labeled data sets, such as ImageNet, have not yet been created in the pharmaceutical sciences. This lack of a standardized data set means that it is difficult to follow existing transfer learning strategies in which one fine-tunes for a new task a model pretrained on a standard data

* *Corresponding author.Wasan, K.M. (Kishor.Wasan@ubc.ca)*


https://doi.org/10.1016/j.drudis.2021.10.022


Please cite this article in press as: A. Bess et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.10.022

set. Hence, it is difficult to transition models trained for discovering drugs for one disease to do the same for another. For AI to be impactful in drug discovery, one needs to develop general techniques and patterns that apply to a range of tasks involving different diseases. In addition, although deep learning[22] has had a central role in the ongoing AI revolution, models developed based on this technique are notorious for their opacity.

Deep neural networks essentially behave like black boxes[23] and do not provide any insight into their underlying decision-making process. This also makes their application in drug discovery onerous. When a drug is flagged by a neural network as being efficacious for a disease, one needs to understand its mechanism of action, the interaction of the drug with the host–protein network, whether the interaction is inhibitory, the pharmacokinetics, the dose–response curve, any associated cytotoxicity, as well as the epistemic and the aleatoric uncertainty associated with the decision of the network. An off-target decision can entail unnecessary costs incurred not only in failed tests *in vitro* and *in vivo*, but also in consequent clinical trials, not to mention the loss of reputation.

## The current pandemic is driving use of AI in drug discovery

Although the above discussion paints a bleak picture of the suitability of AI in drug discovery, there appears to be hope on the horizon. The current COVID-19 pandemic has become the main driving force behind the use of AI to accelerate preclinical drug discovery. At present, a few drugs, such as remdesivir, have been approved by the US Food and Drug Administration (FDA) for off-label use in treating severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) infections. Most of these proposed treatments have been discovered through trial-and-error experiences with the virus by physicians and researchers around the world. It is well documented that the average pharmaceutical company's in-house preclinical discovery cost for a new drug compound is US$209 522 157 (adjusted for inflation) over 3 years (only ~ 12% of all drugs developed eventually get approved by FDA, whereas failed attempts significantly increase the average cost and time requirement of preclinical drug discovery).[24–25] These expenses do not include the costs of basic research at the university level focused on the identification of molecular targets as well as the development of research methods and technologies. The efficiency of drug development, as defined as the successful approval of new pharmaceuticals within the rate of acceptable financial investment, has significantly declined.[25–26] The existing process of creating drugs is slow, inefficient, and costly. Hit identification, lead generation, and lead optimization are key steps at the outset of any drug discovery process. Initially, compounds showing promising activity identified by high-throughput screening as initial hits are filtered and modified to generate lead compounds that satisfy basic drug-likeliness properties.[27] These lead compounds are further optimized to enhance their potency toward the target protein or mechanism as well as to reduce nonselectivity and toxicity. Conventional hit identification is expensive and requires time-consuming screening experiments. Under the circumstances of the current pandemic, the world cannot afford such an inefficient pipeline. What is needed is a principled approach to drug discovery and repurposing that can rapidly address large data sets. This capability will thereby create an improved method for identifying drugs and/or drug combinations that are likely to succeed.

### *Current state of antimicrobial drug discovery*

The enormous time and cost incurred in discovering a new drug compound as well as developing it through the approval process have been so overwhelming in recent times that the pharmaceutical industry has repeatedly shown reduced interest in bringing new drug products to market. The inactivity is most notable in less profitable market segments, such as infectious diseases.[28] Over the past 20 years, the pharmaceutical industry has put infectious disease and antimicrobial drug discovery and development on the backburner. The COVID-19 pandemic has been a distressing reminder of the lack of infrastructure to develop treatments for emerging infectious diseases. The pandemic has been a global reckoning, highlighting the importance of antiviral and antimicrobial drug research for future outbreaks. In recent history, there have been meagre enthusiasm and scarcity of growth in the field of infectious diseases. Case in point, for bacterial infections, every new antibiotic brought to market over the past few decades has only been a slight variation on existing drugs discovered before1984.[29] Only one of the top 50 pharmaceutical companies has antibiotics in clinical development and nearly 75% of the companies currently developing antimicrobials can be regarded as prerevenue, with no approved products in the market.[30–31] Market analysis has shown that drug-resistant forms of these diseases will grow significantly by 2025, with very few new drug strategies in the near future.[32]

## The rise of new AI techniques and their application to drug discovery

Recent advances in AI, with the development of fundamentally new techniques, such as graph neural networks,[33–34] graph embeddings,[35] geometric deep learning,[36] attention networks,[37] self-supervised[38] and unsupervised[39–40] learning, Monte-Carlo graph search,[41] neural networks for protein folding,[42] explainable AI,[43] and generative adversarial networks (GANs),[44] have spurred renewed interest in applying them to accelerate drug discovery. These techniques promise to mitigate the above-mentioned drawbacks of previous-generation AI. They allow for the development of an efficient drug discovery pipeline by leveraging mathematical representations of all interactions between proteins in the host cell.

Using such a model, we can accurately predict whether a particular microbial mechanism will be inhibited by a certain drug. For example, in discovering antivirals, understanding the effects of a drug on viral mechanisms, such as viral entry, RNA transcription, and viral exit, can be crucial for predicting the effectiveness of a therapy involving the drug. Databases, such as HU.MAP,[45] HPIDB,[46] and STRING,[47] provide both human–human and human–virus protein interactions that can be exploited by the above-mentioned techniques. These interactions can be used to provide explanations for why a particular drug compound is efficacious against a disease both in terms of the proteins targeted by the compound and subsequent protein–protein interaction cas-



Please cite this article in press as: A. Bess et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.10.022

cades. For instance, a graph neural network[33–34] can take a graph structure and a feature description for every node as input, to comprehensively model the interactions of a drug within the human interactome, that is, the protein–protein interactions of the human cell. The network learns and operates on the graph structure of the input and ground truth data. Each protein is represented as a node in the graph and the neighborhood of each node is assigned from the set of neighboring nodes in the structure of the protein. Chemical nodes can correspond to existing drugs (including 131 nutraceuticals) in Drugbank, which contains data on 13 580 approved and experimental drugs, or SuperTarget, a large data set of 332 828 drug–target interactions (DTIs). The edges of the graph represent protein interactions. Each protein node could also have features computed from its amino sequence and structure, whereas edges have weights describing interactions experimentally derived between residues. Such a network would be a predominantly encompassing mathematical representation of all physical contacts between proteins within a cell (Fig. 1). ProtVec,[48] a vector representation of protein sequences, would constitute the input features of each protein node. ProtVec is an unsupervised data-driven distributed representation of the protein k-mer sequences as an *n*-dimensional vector in a context-aware manner, useful for neural network predictions or analysis. Target mechanisms would be represented with edges to all proteins associated with them.

The output of such a graph neural network would be node embeddings for each node in the graph. A node embedding characterizes the context of a node with respect to its interaction with other nodes in the graph. Fig. 2 visualizes the embeddings of such a graph in 2D using t-SNE.[49] The red clusters in Fig. 2 show how drugs are clustered, whereas blue clusters show the clustering of the proteins. Overlap of the blue clusters with the red clusters indicates drug–protein interactions.

The DeepDrug team (see below) developed such node embeddings to be inputted into a Siamese network.[50] Siamese networks project embeddings into multidimensional space and calculate distance between them within that dimensionality. The closer the prediction is to zero; the higher the interaction between a pair of embeddings. Such a Siamese network will take embeddings of a pair of drug–protein nodes as input. The network would output a distance metric indicating the effect of the drug on target proteins and viral mechanisms involving them. For example, for the nutraceutical biotin, the Siamese network predicts Abelson kinase (ABL1) as a target protein. It is known from the literature[51] that Abelson kinase inhibitors can have effectiveness against SARS and Middle East respiratory syndrome (MERS) coronavirus infections. Similarly, the nutraceutical levomenol, a chamomile extract, is predicted to target signal transducer and activator of transcription 3 (STAT3). The literature[52] shows that inhibition of STAT3 can help regulate cytokine storms that might result in acute respiratory distress syndrome (ARDS) during COVID-19 infection. One could use a Bayesian Siamese network[53] with weights sampled from a Gaussian distribution to further provide uncertainty estimates for its predictions. Geometric deep-learning techniques can also generalize such graph neural networks and can efficiently extract representations of chemical features.[54]

The resulting weights with their uncertainty estimates can be used to prioritize drugs and filter the top drug candidates by taking their respective toxicities and synthetic accessibilities[24] into consideration using a multicriterion optimization algorithm. This multicriteria optimization algorithm can: (i) rank all FDA-approved drugs according to the weight/uncertainty estimates as obtained from the Siamese network; and (ii) solve an optimization problem that will shortlist drugs with the highest weight/certainty, lowest toxicity score, and highest synthetic accessibility score.

Another important advance in AI that can significantly impact drug discovery is explainable AI (XAI). Confidence-aware networks[55–57] have helped lift the veil on the opaque decision-making process of deep neural networks. It is now possible to understand the epistemic and aleatoric uncertainty associated with the decision of a deep neural network. Indeed; when a confidence-aware neural network predicts that a drug is efficacious against a particular disease, it will also provide a measure of its own confidence in its prediction. High confidence predictions can proceed for *in vitro* validation, whereas low confidence predictions can be filtered out. Recent advances in transfer learning[58–60] also bode well for drug discovery. Domain adaptation[61] now allows models trained to predict drugs targeting one disease to be reused to recommend those for another. Transfer learning together with low-shot techniques[62–63] alleviate the need for large; labeled data sets in model training. Currently; tasks such as predicting toxicity or drug–drug interactions (DDIs) require large volumes of labeled training data. Acquiring such data in the pharmaceutical vertical is difficult because labeling requires domain knowledge. This markedly hinders the development of essential tools for drug discovery. Modern unsupervised[39–40] and self-supervised learning techniques[38] can ease the problem by exploiting vast amounts of available unlabeled data.

## Renewed efforts in applying AI to drug discovery

Stunning advances in AI, as described above, have spurred renewed interest in using AI to accelerate preclinical drug discovery. Several teams have been working with AI platforms to repurpose existing drugs and re-engineer new drugs in the pursuit of finding life-saving medicines. Here, we highlight platforms with state-of-the-art machine learning and AI technology that are spearheading new methods for drug discovery. Recently, Bender and Cortés-Ciriano published a paper discussing whether AI was having an impact on drug discovery and limitations of this approach to date.[64–65] Here, we address the concerns raised by these authors and provide a brief introduction to the implementation, strategy, and successes of each team. Each of these methods can lead to both theoretical and practical applications in drug discovery.

### *BenevolentAI*

The BenevolentAI team is working on a drug discovery approach that involves the use of biological knowledge graphs to identify new treatments.[66] Using an AI technique called natural language processing (NLP),[67–68] knowledge graphs are extracted from the scientific literature to identify previously unknown correlations.[69] The resulting graph represents an interlinked network



Please cite this article in press as: A. Bess et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.10.022

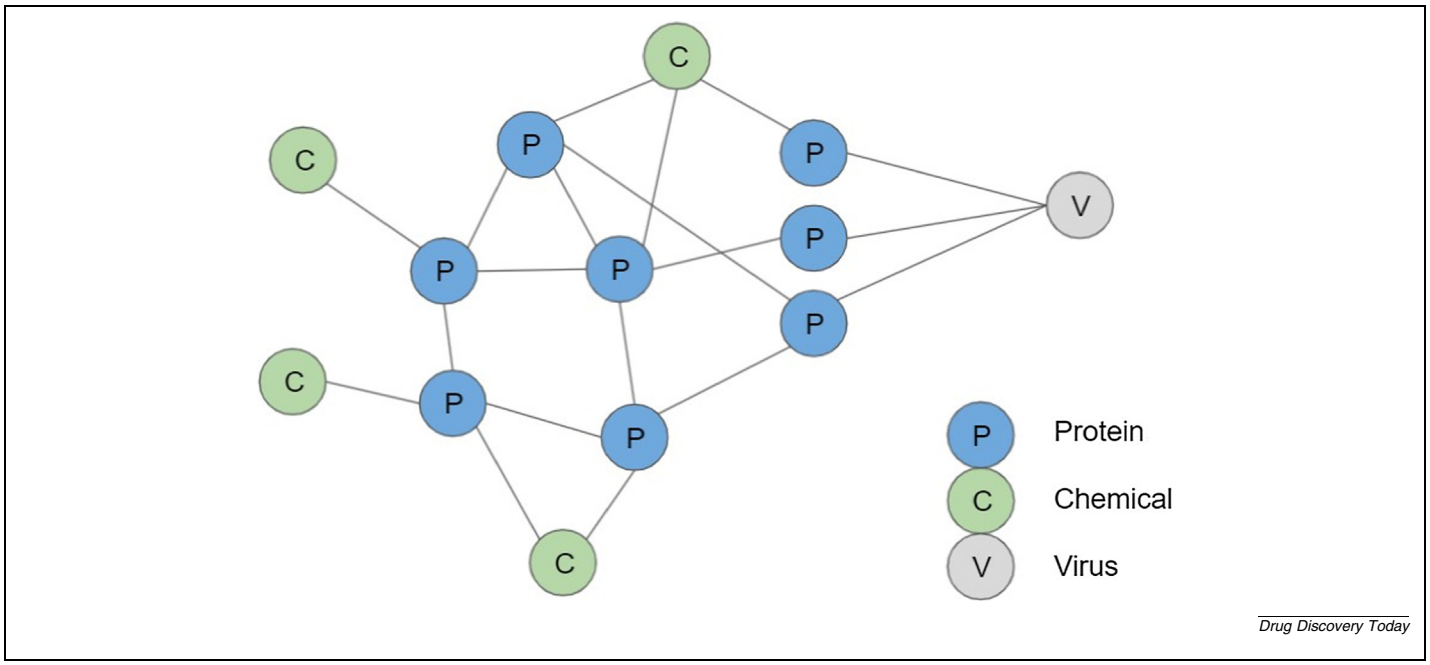


**FIGURE 1**

Visualization of protein–protein and protein–chemical graphs. The blue dots represent protein nodes, the green dots represent chemical nodes, the gray dot represents a virus protein, and the lines represent edges in the graph (protein–protein or chemical–protein interactions).

of concepts that places scientific data in context by linking semantic metadata. This framework allows the BenevolentAI team to integrate previously unconnected research to identify links that could be targets for drug development. This network was used to identify baricitinib;[70] a drug approved for the treatment of rheumatoid arthritis; as a repurposed treatment for COVID-19 in mitigating the cytokine storm through inhibition of adaptor-associated protein kinase 1 (AAK1). By making use of this knowledge base, the team was able to complete this analysis by February 2020, only weeks after the first COVID-19 case was reported in the USA. By November of the same year, BenevolentAI and Eli Lilly had completed clinical trials and received an Emergency Use Authorization from the FDA as a treatment for COVID-19.

BenevolentAI also has a secondary project[71] to analyze and compare 3D binding sites in which both positive and negative binding pairs of protein-pockets and ligands are used to train a network for protein-pocket matching. By encoding the 3D shapes of the binding sites, BenevolentAI's network is able to learn which features of a protein-pocket representation predict binding affinity and can screen many pockets to identify novel drug targets. This machine learning approach is called 'field of distance metric learning',[72] and enables the BenevolentAI team to predict results of previously unknown DTIs.

### *Atomwise*

Another emerging platform is Atomwise, which uses an AI technique called convolutional neural networks (CNNs) to analyze the biological activity to predict the binding affinity of small molecules.[73] CNNs are a class of neural networks mainly used to understand imagery. Molecular shape analysis of small molecules using CNNs can predict binding affinity measurements of different molecules to protein structures. This allows Atomwise to predict the biological activity and pharmacology of small molecules for drug discovery. The Atomwise networks apply feature locality and hierarchical composition to model pharmacological activity and chemical interactions. Their networks showed promising results for the Database of Useful (Docking) Decoys Enhanced (DUD-E), achieving an area under the curve (AUC) greater than 0.9 on 57.8% of the docking targets in DUD-E.[21,73] Atomwise used this technology to screen millions of molecules against known SARS-CoV-2 proteins to explore broad-spectrum therapies for the treatment of COVID-19 and other coronavirus infections.

### *Insilico medicine*

The Insilico Medicine team proposed a unique generative adversarial network (GAN)-based approach for synthesizing new drugs for individual diseases.[74] GANs function by discovering patterns in input data from which the model can generate new samples that could have plausibly been drawn from the original data set. Insilico Medicine's GAN network synthesizes new compounds by iteratively generating molecules; while analyzing certain molecular parameters, such as biological activity and synthetic feasibility. The system then optimizes across its set parameters and generates new molecules until it reaches a local maximum. Such a network can generate molecules with certain properties or activities against a pharmacological target, making the network useful for initial discovery. However, only a few examples of generative drug design have achieved validation in *in vitro* or *in vivo* experiments.

Insilico Medicine originally focused their efforts on generating chemotypes targeting the SARS-CoV-2 main protease. By 4 February, 2020, Insilico Medicine released their first potential



Please cite this article in press as: A. Bess et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.10.022

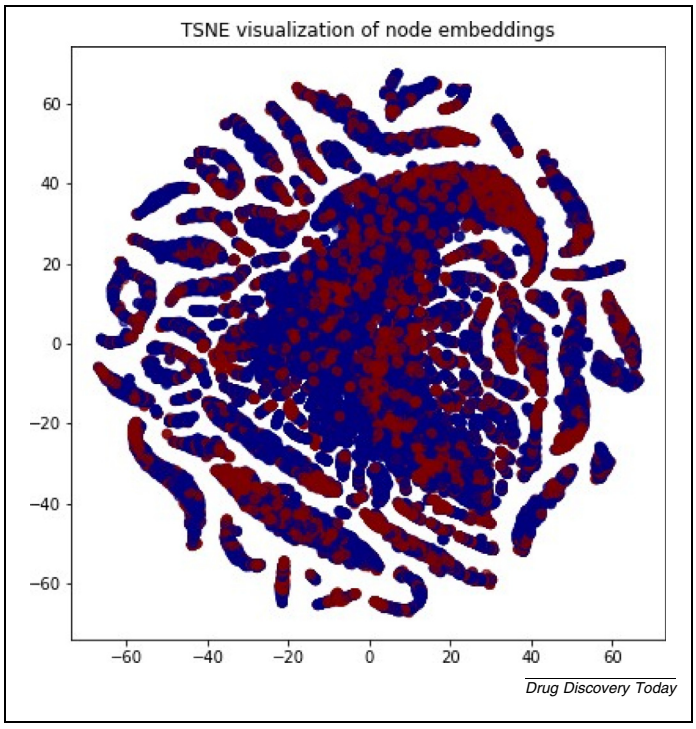


**FIGURE 2**

Visualization of node embeddings in two dimensions using tSNE. The red clusters show how the drugs are clustered, whereas the blue clusters show the clustering of the proteins. Overlap of the blue clusters with the red clusters represent drug–protein interactions.

*de novo* protease inhibitor. The Insilico Medicine team recently published ten representative structures of protease inhibitors for potential development against COVID-19.[75] Even so, the greatest complication of using a GAN lies in the nature of the network itself. Any output from such a GAN is derived within a 'black box' system, giving researchers little to no explanation or understanding of the underlying analyses. Given that the patterns and regularities identified in the data are known only by the AI system, extensive laboratory testing is required to confirm any findings from this technique.

### *ComboNet*

The ComboNet team at the Broad Institute (Cambridge, MA, USA) leveraged DTIs to identify synergistic combinations against SARS-CoV-2.[76] The ComboNet system predicts DTIs from the molecular structures of the compounds analyzed. The ComboNet architecture comprises two major components: a graph convolutional network (GCN), which is trained to represent the molecular structure of the compound, and a model for target–disease association. The advantage of using this methodology is the ability to predict from compounds with incomplete DTI information. The second model learns how biological targets and molecular structure features interact to present antiviral activity and synergy. The team used training data from NIH's NCATS cytopathic effect assay against SARS-CoV-2 as well as SARS-CoV-2 drug combination assays with synergy scored using the BLISS model.[77]

### *DeepDrug*

The DeepDrug team, a semifinalist in the IBM Watson Artificial Intelligence XPRIZE competition, created an efficient AI-based platform to design new compounds and repurpose existing drugs for emerging infectious diseases.[24,26,78] The DeepDrug pipeline is capable of automatically synthesizing targeted drug molecules using beam search techniques,[79] as well as filtering candidates based on chemical criteria (e.g., Lipinski's Rule of Five)[27 ]and potential adverse effects. This allows the team to predict the candidates that are most likely to succeed in the patient population. The pipeline is modular in nature and currently comprises eMolFrag,[26] eSynth,[78] eToxPred,[24] eDrugRes, eVir, and several other AI-based filters. Given a collection of molecules, eMolFrag generates a set of unique fragments and pharmacophores that act as 'building blocks'. Fig. 3 shows the ability of eMolFrag to identify bioactive building blocks from known drugs. eSynth[78] uses beam search techniques[79] to combine these molecular fragments into novel molecules *de novo*. It assembles millions of molecules in minutes, while logging the associated chemical reactions used to construct each molecule. This trace of chemical reactions can be used to synthesize any of these molecules in a wet lab setting. These molecules can be then further filtered for toxicity, specificity, and ease of manufacturing.

Using two of these modules, the DeepDrug team synthesized an adenosine receptor from components acquired by decomposing four adenosine receptor antagonists.[26] Adenosine receptor antagonists have important roles in inflammation, pain, and immune responses, making them attractive targets for pharmacotherapy.

eToxPred,[24] the third module in the DeepDrug pipeline, is used to estimate toxicity and synthetic accessibility of small molecules. Estimating toxicity is a key component of the overall DeepDrug pipeline, to rapidly and proactively filter out compounds with undesirable or adverse effects. In contrast to other approaches that use manually crafted descriptors,[80] eToxPred uses the molecular fingerprints of the chemical compounds to model toxicity directly, making it more effective against highly diverse data sets. Fig. 4 shows eToxPred using machine-learning techniques to filter the candidate drug molecules with respect to their potential toxicity based on structural properties. The output eToxPred value is a Tox-score between zero and one, with zero being the least toxic and one indicating a high likelihood for toxicity. FDA-approved drugs have the lowest median Tox-score of 0.34, whereas the toxicity of active compounds from the DUD-E data set is slightly higher, with a median Tox-score of 0.46. Molecules in both natural product and traditional herbal medicine data sets show higher toxicity scores with a median Tox-scores of $\sim$0.55. These results are validated by other studies that examine the potentially toxic constituents, which include alkaloids, glycosides, polypeptides, amino acids, phenols, organic acids, terpenes, and lactones.

eDrugRes was created to identify effective chemicals against antibiotic-resistant bacteria by exploring drug effects and mutations within microbial protein–protein interaction networks. This system uses GCNs to predict whether a specific chemical compound would have therapeutic activity against certain strains of bacteria.





Please cite this article in press as: A. Bess et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.10.022

Several new modules have recently been added to the DeepDrug pipeline. The first is eVir, which can determine viral specificity of drugs with the goal of repurposing existing drugs. It uses an AI technique to generate a fingerprint for drugs and known antiviral peptides (AVPs)[81] that captures their properties and context within a mathematical representation of all cellular protein interactions. By comparing these fingerprints in the context of the data, the system provides separate predictions for three mechanisms of viral infection (e.g., entry, fusion, and replication), which affords a higher degree of specificity in drug selection. This enables eVir to explain its predictions based on specific correlated mechanisms and protein interactions. The DeepDrug team has used eVir to identify multiple drugs and drug therapies with high likelihood of efficacy against SARS-CoV-2. These therapies have demonstrated their effectiveness against SARS-CoV-2 infection, both in *in vitro* studies (with Vero E6[82] and Calu-3 cells[83] as well as *in vivo* studies using transgenic mice. Finally, the DeepDrug AI platform can predict the DDIs in drug combinations as well as the synergy of specific drug combination therapies with the latest module, eComb. Recently, an oral drug combination therapy for COVID-19, discovered by the DeepDrug AI platform, started human trials at the Riverside University Health System, California, and in Europe.[84] In addition, based on the AI analysis above, the nutraceuticals biotin and levomenol were identified to have potential effects against SARS-CoV. The DeepDrug team combined these two nutraceuticals with other essential vitamins and minerals to create a dietary supplement known as Inhibinol.[85]

## Comparison of technologies

Drug discovery is associated with complex workflows that have multiple aspects spanning. The above-mentioned innovative teams (Table 1) are each working on specific verticals pertinent to drug discovery. Depending on their particular use cases, the teams use diverse techniques, each of which has their own advantages and disadvantages. For instance, the Insilico Medicine team uses GANs, the underlying analysis of which is difficult to explain. However, when applied in the context of COVID-19, the team identified ten proteasomal inhibitors that are currently being testing in labs by several research groups worldwide. Unlike Insilico Medicine, Atomwise's system is only capable of repurposing known molecules. However, their approach requires a large volume of experimental and structural data. By contrast, BenevolentAI leveraged a massive data set and previously developed knowledge graphs to become the first team to identify a possible inhibitor for cytokine storms: baricitinib. The disadvantage of BenevolentAI's system is its limited capability in discovering known molecules based only on natural language processing from a corpus of existing literature. BenevolentAI also has protein-binding prediction networks still in early phase testing. ComboNet is designed to predict drug synergy by modeling compound and biological target structural features with a GCN. The advantage of this technique is the ability to predict DTIs for compounds with incomplete experimental data. The disadvantage is that the structural training set is hyperspecific to a few key viral SARS-CoV-2 proteins, whereas the drug combinations are based on old curated data with previously tested drugs, such as remde-

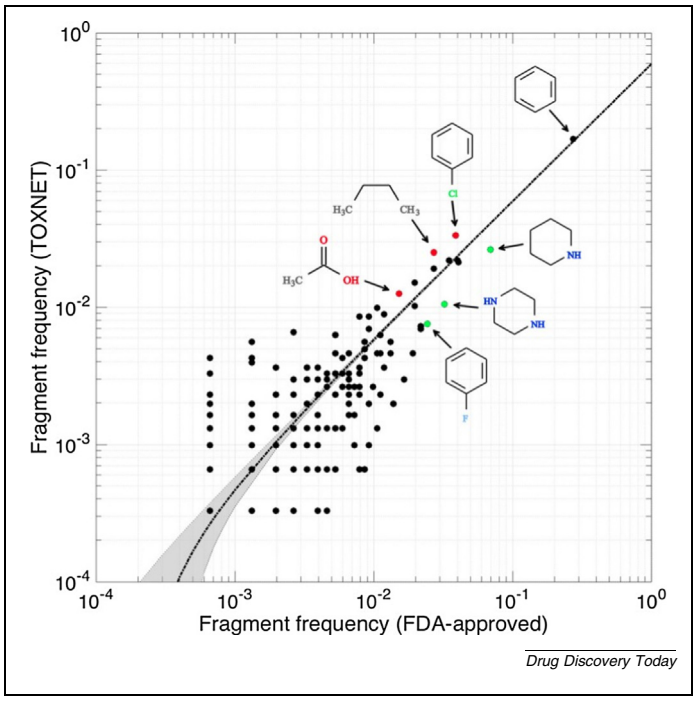


**FIGURE 3**

Composition of nontoxic and toxic compounds. The scatter plot shows the frequencies of eMolFrag-extracted chemical fragments from US Food and Drug Administration (FDA)-approved (nontoxic) and TOXNET (toxic) molecules. The dotted black line is the line of regression, and the gray area represents the corresponding confidence intervals. Examples of three commonly found FDA-approved fragments (piperidine, piperazine, and fluorophenyl) are in green, whereas fragments of more commonly toxic fragments from the TOXNET data set (chlorophenyl, n-butyl, and acetic acid) are in red. Adapted from.[24]

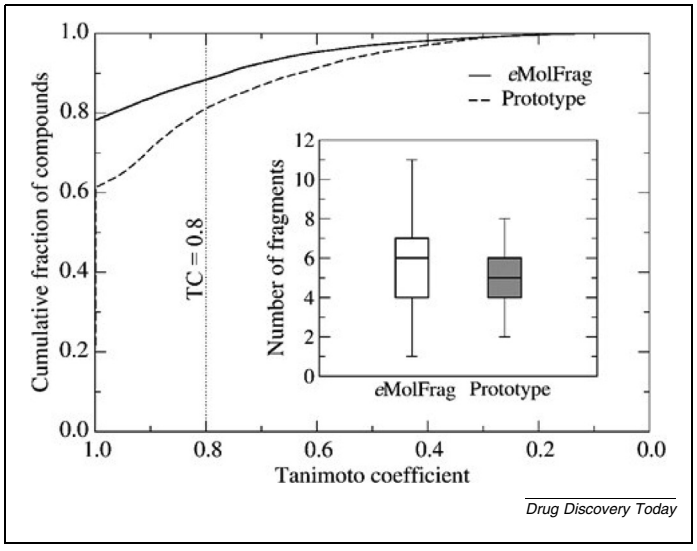


**FIGURE 4**

Bioactive compounds from the Database of Useful (Docking) Decoys Enhanced (DUD-E) database fragmented with eMolFrag. eMolfrag was able to generate an average of six fragments per molecule. eSynth uses beam search techniques to create new drug molecules by combining the building blocks generated by eMolFrag in a chemically comprehensive way. By using fragments generated by eMolFrag, eSynth could reconstruct 78.3% of active compounds with a Tanimoto coefficient (TC) of 1.0 and 88.4% with a TC $\geq$ 0.8. Adapted from.[26]



Please cite this article in press as: A. Bess et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.10.022

**TABLE 1**

**Comparison of emerging AI teams and their respective technologies.**

| AI team | Technique | *In vitro* | *In vivo* | Clinical trials |
|---|---|---|---|---|
| BenevolentAI | Knowledge graphs and protein pocket analysis | ✓ | ✓ | ✓ |
| Atomwise | Molecular docking prediction; GAN *de novo* synthesis of chemotypes | ✓ | | |
| Insilico Medicine, ComboNet | GCN *in silico* analysis of drug combinations | ✓ | | |
| DeepDrug | eMolFrag, eSynth, eToxPred, eDrugRes, eVir, eComb | ✓ | ✓ | ✓ (Approved May 2021 |

sivir. Without extensive testing on a disjoint test set, it is unclear whether such a training set would be able to predict accurately for compounds outside the training set. Unfortunately, ComboNet have only tested their predicted combination therapies for SARS-CoV-2 against Vero E6 cells *in vitro*. Finally, DeepDrug is capable of both synthesizing new molecules *de novo* or repurposing drugs, while predicting their likelihood of human toxicity, manufacturing difficulty, and target specificity.

Overall, AI in drug discovery is an extremely powerful but nascent tool. Companies and teams have designed systems that handle only specific types of analyses proficiently. Since each team's respective data sets are meticulously aggregated and collated individually, their frame of reference might only be useful in a narrow vertical. Additionally, such data are considered proprietary and are often siloed within the team. For instance, the recommendations provided by the existing AI pipelines do not consider the pre-existing conditions of patients. Such global contextual information could be provided in the form of deidentified patient electronic health records. Access to such data would allow for more context-sensitive recommendations that can be valuable in a clinical setting. Overall, these emerging AI tools can be utilized to move the ball toward an ultimate goal: rapidly identifying treatments for infectious diseases. Although certain types of analysis, such as drug combination synergy, expected dosage, and adverse drug reactions, are also important, predictive algorithms for these aspects have yet to be extensively developed. From toxicology to DDIs, to drug–protein specificity, scientists are trying to perfect these prediction systems in every aspect of drug discovery. In the long run, these technologies are a first step toward a comprehensive pipeline capable of rapidly identifying key drugs to combat any emerging infectious disease at a fraction of the time and cost.

## Concluding remarks and outlook

The current drug development process is slow, inefficient, and costly. There is a dire need to develop new platforms and approaches to combat diseases quickly compared with traditional approaches. AI applications in other sectors are massively improving platform efficiencies, refining targeted results, and transforming labor-intensive processes. Such efficiencies are key in disrupting the current stagnation of the pharmaceutical industry. Big pharma's insufficient response to emerging pathogens burdens healthcare systems across the globe and ultimately costs lives. Data projection, mining, and analysis at scale will assist scientists and pharmacologists in identifying the most effective compounds by cross-checking millions of chemical combinations. All of the AI platforms described in this paper are applying cutting-edge techniques to their respective complex pharmacological challenges. These novel approaches for drug discovery and development are a transformative first step in disruption of the pharmaceutical industry. We need to embrace these new technologies and strategies amid the turmoil of the current COVID-19 pandemic.

## References

1 S. Mukhopadhyay, S. Iyengar, A.M. Madni, R. Di Biano, The next generation of artificial intelligence: synthesizable AI, Advances in Intelligent Systems and Computing 880 (2018) 659–677.
2 M. Daily, S. Medasani, R. Behringer, M. Trivedi, Self-driving cars, Computer 50 (2017) 18–23.
3 E. Gibney, Google AI algorithm masters ancient game of Go, Nature News 529 (2016) 445.
4 S. Risi, M. Preuss, From chess and Atari to Starcraft and beyond: how game AI is driving the world of AI, KI-Künstliche Intelligenz 34 (2020) 7–17.
5 K. Gruber, Is the future of medical diagnosis in computer algorithms?, Lancet Digital Health 1 (2019) e15–e16
6 F. Shamsi, B. Aneja, P. Hasan, et al., Synthesis, anticancer evaluation and DNA-binding spectroscopic insights of quinoline-based 1,3,4-oxadiazole-1,2,3-triazole conjugates, ChemistrySelect 4 (41) (2019) 12176–12182.
7 Liu Q, Mukhopadhyay S, Rodriguez MXB. A one-shot learning framework for assessment of fibrillar collagen from second harmonic generation images of an infarcted myocardium. *arXiv* 2020; 2020: arXiv:2001.08395v2
8 S.S. Iyengar, X. Li, H. Xu, S. Mukhopadhyay, N. Balakrishnan, A. Sawant, et al., Toward more precise radiotherapy treatment of lung tumors, Computer 45 (1) (2012) 59–65.
9 Basu S, Ganguly S, Mukhopadhyay S, DiBiano R, Karki M, Nemani R. Deepsat: a learning framework for satellite imagery. In: *SIGSPATIAL '15: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. New York; Association for Computing Machinery; 2015: 37
10 Q. Liu, S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, et al., Deepsat v2: feature augmented convolutional neural nets for satellite image classification, Remote Sensing Letters 11 (2020) 156–165.
11 S. Basu, S. Ganguly, R.R. Nemani, S. Mukhopadhyay, G. Zhang, C. Milesi, et al., A semiautomated probabilistic framework for tree-cover delineation from 1-m NAIP imagery using a high-performance computing architecture, IEEE Transactions on Geoscience and Remote Sensing 53 (10) (2015) 5690–5708.
12 C. Chokwitthaya, Y. Zhu, S. Mukhopadhyay, E. Collier, Augmenting building performance predictions during design using generative adversarial networks and immersive virtual environments, Automation in Construction 119 (2020) 103350.
13 C. Chokwitthaya, Y. Zhu, R. Dibiano, S. Mukhopadhyay, Combining context-aware design-specific data and building performance models to improve building performance predictions during design, Automation in Construction 107 (2019) 102917.
14 Nabijiang A, Mukhopadhyay S, Zhu Y, Gudishala R, Saeidi S, Liu Q. Why do you take that route? *arXiv* 2019; 2019: arXiv:1905.06463.
15 F.G. Ferreira, A.H. Gandomi, R.T. Cardoso, Artificial intelligence applied to stock market trading: a review, IEEE Access 9 (2021) 30898–30917.
16 Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: *IEEE Conference on Computer Vision and Pattern Recognition*. New York; IEEE: 2009: 248–255.



Please cite this article in press as: A. Bess et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.10.022

17 D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, et al., DrugBank 5.0: a major update to the DrugBank database for 2018, Nucleic Acids Research 46 (D1) (2018) D1074–D1082.
18 M.K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, J. Chong, BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology, Nucleic Acids Research 44 (D1) (2016) D1045–D1053.
19 M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, M. Tanabe, KEGG: integrating viruses and cellular organisms, Nucleic Acids Research 49 (D1) (2021) D545–D551.
20 S. Günther, M. Kuhn, M. Dunkel, M. Campillos, C. Senger, E. Petsalaki, et al., SuperTarget and Matador: resources for exploring drug-target relationships, Nucleic Acids Research 36 (Suppl. 1) (2007) D919–D922.
21 M.M. Mysinger, M. Carchia, J.J. Irwin, B.K. Shoichet, Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking, Journal of Medicinal Chemistry 55 (14) (2012) 6582–6594.
22 Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
23 Marcus G. Deep learning: a critical appraisal. *arXiv* 2018; 2018: arXiv:1801.00631.
24 L. Pu, M. Naderi, T. Liu, H.C. Wu, S. Mukhopadhyay, M. Brylinski, eToxpred: a machine learning-based approach to estimate the toxicity of drug candidates, BMC Pharmacology and Toxicology 20 (1) (2019) 1–15.
25 S.M. Paul, D.S. Mytelka, C.T. Dunwiddie, C.C. Persinger, B.H. Munos, S.R. Lindborg, et al., How to improve R&D productivity: the pharmaceutical industry's grand challenge, Nature Reviews Drug Discovery 9 (3) (2010) 203–214.
26 T. Liu, M. Naderi, C. Alvin, S. Mukhopadhyay, M. Brylinski, Break down in order to build up: decomposing small molecules for fragment-based drug design with eMolFrag, Journal of Chemical Information and Modeling 57 (4) (2017) 627–631.
27 C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Advanced Drug Delivery Reviews 23 (1–3) (1997) 3–25.
28 P. Trouiller, P. Olliaro, E. Torreele, J. Orbinski, R. Laing, N. Ford, Drug development for neglected diseases: a deficient market and a public-health policy failure, Lancet 359 (9324) (2002) 2188–2194.
29 Pew. *Tracking the Global Pipeline of Antibiotics in Development, April 2020*. www.pewtrusts.org/en/research-and-analysis/issue-briefs/2020/04/tracking-the-global-pipeline-of-antibiotics-in-development [Accessed 28 October 2021].
30 B. Plackett, Why big pharma has abandoned antibiotics, Nature 586 (7830) (2020) S50–S52.
31 Pew. *A Scientific Roadmap for Antibiotic Discovery*. www.pewtrusts.org/en/research-and-analysis/reports/2016/05/a-scientific-roadmap-for-antibiotic-discovery [Accessed 28 October 2021].
32 WHO. Lack of new antibiotics threatens global efforts to contain drug-resistant infections. www.who.int/news/item/17-01-2020-lack-of-new-antibiotics-threatens-global-efforts-to-contain-drug-resistant-infections [Accessed 28 October 2021].
33 Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, S.Y. Philip, A comprehensive survey on graph neural networks, IEEE Transactions on Neural Networks and Learning Systems 32 (1) (2020) 4–24.
34 Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, Gómez-Bombarelli R, Hirzel T, Aspuru-Guzik A, *et al*. Convolutional networks on graphs for learning molecular fingerprints. *arXiv* 2015; 2015: arXiv:1509.09292.
35 Yang Z, Cohen W, Salakhudinov R. Revisiting semi-supervised learning with graph embeddings. *arXiv* 2016; 2016: arXiv:1603.08861v2.
36 M.M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, P. Vandergheynst, Geometric deep learning: going beyond euclidean data, IEEE Signal Processing Magazine 34 (4) (2017) 18–42.
37 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, *et al*. Attention is all you need. *arXiv* 2017; 2017: arXiv:1706.03762v5.
38 Jing L, Tian Y. Self-supervised visual feature learning with deep neural networks: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2020; XX: XXX–YYY.
39 Liu Q, Mukhopadhyay S. Unsupervised learning using pretrained CNN and associative memory bank. *arXiv* 2018; 2018: arXiv:1805.01033v1.
40 Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. *arXiv* 2020; 2020: arXiv:2002.05709v3.
41 D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. van den Driessche, et al., Mastering the game of Go with deep neural networks and tree search, Nature 529 (7587) (2016) 484–489.
42 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, et al., Highly accurate protein structure prediction with AlphaFold, Nature 596 (7873) (2021) 583–589.
43 D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.Z. Yang, XAI: explainable artificial intelligence. *Science*, Robotics 4 (37) (2019) XXX.
44 I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., Generative adversarial networks, Communications of the ACM 63 (11) (2020) 139–144.
45 K. Drew, C. Lee, R.L. Huizar, F. Tu, B. Borgeson, C.D. McWhite, et al., Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes, Molecular Systems Biology 13 (6) (2017) 932.
46 M.G. Ammari, C.R. Gresham, F.M. McCarthy, B. Nanduri, HPIDB 2.0: a curated database for host–pathogen interactions, Database 2016 (2016) baw103.
47 D. Szklarczyk, A.L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, et al., STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets, Nucleic Acids Research 47 (D1) (2019) D607–D613.
48 E. Asgari, M.R. Mofrad, Continuous distributed representation of biological sequences for deep proteomics and genomics, PLoS ONE 10 (11) (2015) e0141287.
49 V. Maaten, dL, Hinton G., Visualizing data using t-SNE, Journal of Machine Learning Research 9 (11) (2008) XXX.
50 Bertinetto L, Valmadre J, Henriques JF, Vedaldi A, Torr PH. Fully-convolutional Siamese networks for object tracking. *arXiv* 2016; 2016: arXiv:1606.09549v2.
51 C.M. Coleman, J.M. Sisk, R.M. Mingo, E.A. Nelson, J.M. White, M.B. Frieman, Abelson kinase inhibitors are potent inhibitors of severe acute respiratory syndrome coronavirus and Middle East respiratory syndrome coronavirus fusion, Journal of Virology 90 (19) (2016) 8924–8933.
52 A. Jafarzadeh, M. Nemati, S. Jafarzadeh, Contribution of STAT3 to the pathogenesis of COVID-19, Microbial Pathogenesis 154 (2021) 104836.
53 H. Wang, D.Y. Yeung, Towards Bayesian deep learning: a framework and some existing methods, IEEE Transactions on Knowledge and Data Engineering 28 (12) (2016) 3395–3408.
54 P. Hop, B. Allgood, J. Yu, Geometric deep learning autonomously learns chemical features that outperform those engineered by domain experts, Molecular Pharmaceutics 15 (10) (2018) 4371–4377.
55 Moon J, Kim J, Shin Y, Hwang S. Confidence-aware learning for deep neural networks. *arXiv* 2020; 2020: arXiv:2007.01458v3.
56 DeVries T, Taylor GW. Learning confidence for out-of-distribution detection in neural networks. *arXiv* 2018; 2018: arXiv:1802.04865.
57 Lakshminarayanan B, Tran D, Liu J, Padhy S, Bedrax-Weiss T, Lin Z. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *arXiv* 2020; 2020: arXiv:2006.10108v2.
58 Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks?. *arXiv* 2014; 2014: arXiv:1411.1792.
59 E. Collier, R. DiBiano, S. Mukhopadhyay, Cactusnets: layer applicability as a metric for transfer learning, in: 2018 International Joint Conference on Neural Networks. IEEE, 2018, pp. 1–8.
60 E. Collier, S.G.A.P. Mukhopadhyay, quantifying the generative adversarial set and class feature applicability of deep neural networks, in: 2020 25th International Conference on Pattern Recognition. IEEE, 2021, pp. 8384–8391.
61 Long M, Zhu H, Wang J, Jordan MI. Unsupervised domain adaptation with residual transfer networks. *arXiv* 2016; 2016: arXiv:1602.04433 2016.
62 Wang YX, Girshick R, Hebert M, Hariharan B. Low-shot learning from imaginary data. *arXiv* 2018; 2018: arXiv:1801.05401v2.
63 E. Collier, S. Mukhopadhyay, K. Duffy, S. Ganguly, G. Madanguit, S. Kalia, et al., Semantic segmentation of high resolution satellite imagery using generative adversarial networks with progressive growing, Remote Sensing Letters 12 (5) (2021) 439–448.
64 A. Bender, I. Cortes-Ciriano, Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: ways to make an impact, and why we are not there yet, Drug Discovery Today 26 (2) (2021) 511–524.
65 A. Bender, I. Cortes-Ciriano, Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data used for AI in drug discovery, Drug Discovery Today 26 (4) (2021) 1040–1052.
66 J. Stebbing, A. Phelan, I. Griffin, C. Tucker, O. Oechsle, D. Smith, et al., COVID-19: combining antiviral and anti-inflammatory treatments, Lancet Infectious Diseases 20 (4) (2020) 400–402.
67 C. Manning, H. Schutze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, 1999.
68 Wolf T, Chaumond J, Debut L, Chaumond J, Delangue C, Moi A, *et al*. Transformers: state-of-the-art natural language processing. In: *Proceedings of the 2020 EMNLP (Systems Demonstration)*. Stroudsburg; Association for Computational Linguistics; 2020: 38–45.



Please cite this article in press as: A. Bess et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.10.022

69 Fauqueur J, Thillaisundaram A, Togia T. Constructing large scale biomedical knowledge bases from scratch with rapid annotation of interpretable patterns. *arXiv* 2019; 2019: arXiv:1907.01417.

70 P. Richardson, I. Griffin, C. Tucker, D. Smith, O. Oechsle, A. Phelan, et al., Baricitinib as potential treatment for 2019-nCoV acute respiratory disease, Lancet 395 (10223) (2020) e30.

71 M. Simonovsky, J. Meyers, DeeplyTough: learning structural comparison of protein binding sites, Journal of Chemical Information and Modeling 60 (4) (2020) 2356–2366.

72 L. Yang, R. Jin, Thesis Distance metric learning: a comprehensive survey, Michigan State University 2 (2006) 4.

73 Wallach I, Dzamba M, Heifets A. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv* 2015; 2015: arXiv:1510.02855.

74 A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper, A. Zhavoronkov, druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico, Molecular Pharmaceutics 14 (9) (2017) 3098–3104.

75 Zhavoronkov A, Aladinskiy V, Zhebrak A, Zagribelnyy B, Terentiev V, Bezrukov DS, *et al*. Potential 2019-nCoV 3C-like protease inhibitors designed using generative deep learning approaches. *ChemRxiv*. Published online February 19, 2020. http://dx.doi.org/10.26434/chemrxiv.11829102.v2.

76 Akal O, Peng Z, Valadez GH. ComboNet: combined 2D & 3D architecture for aorta segmentation. *arXiv* 2020; 2020: arXiv:2006.05325 2020.

77 T. Bobrowski, L. Chen, R.T. Eastman, Z. Itkin, P. Shinn, C.Z. Chen, et al., Synergistic and antagonistic drug combinations against SARS-CoV-2, Molecular Therapy 29 (2) (2021) 873–885.

78 M. Naderi, C. Alvin, Y. Ding, S. Mukhopadhyay, M. Brylinski, A graph-based approach to construct target- focused libraries for virtual screening, Journal of Cheminformatics 8 (1) (2016) 1–16.

79 A. Kumar, S. Vembu, A.K. Menon, C. Elkan, Beam search algorithms for multilabel learning, Machine Learning 92 (1) (2013) 65–89.

80 A. Mayr, G. Klambauer, T. Unterthiner, S. Hochreiter, DeepTox: toxicity prediction using deep learning, Frontiers in Environmental Science 3 (2016) 80.

81 N. Thakur, A. Qureshi, M. Kumar, AVPpred: collection and prediction of highly effective antiviral pep- tides, Nucleic Acids Research 40 (W1) (2012) W199–W204.

82 M.L. Ng, S.H. Tan, E.E. See, E.E. Ooi, A.E. Ling, Proliferative growth of SARS coronavirus in Vero E6 cells, Journal of General Virology 84 (12) (2003) 3291–3303.

83 K.A. Foster, M.L. Avery, M. Yazdanian, K.L. Audus, Characterization of the Calu-3 cell line as a tool to screen pulmonary drug delivery, International Journal of Pharmaceutics 208 (1–2) (2000) 1–11.

84 PRNewswire. Human studies begin on artificial Intelligence discovered COVID-19 treatment with up to 97 percent effectiveness. https://finance.yahoo.com/news/human-studies-begin-artificial-intelligence-130000945.html [Accessed 28 October 2021].

85 https://inhibinol.com/ [Accessed 28 October 2021].





Please cite this article in press as: A. Bess et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.10.022